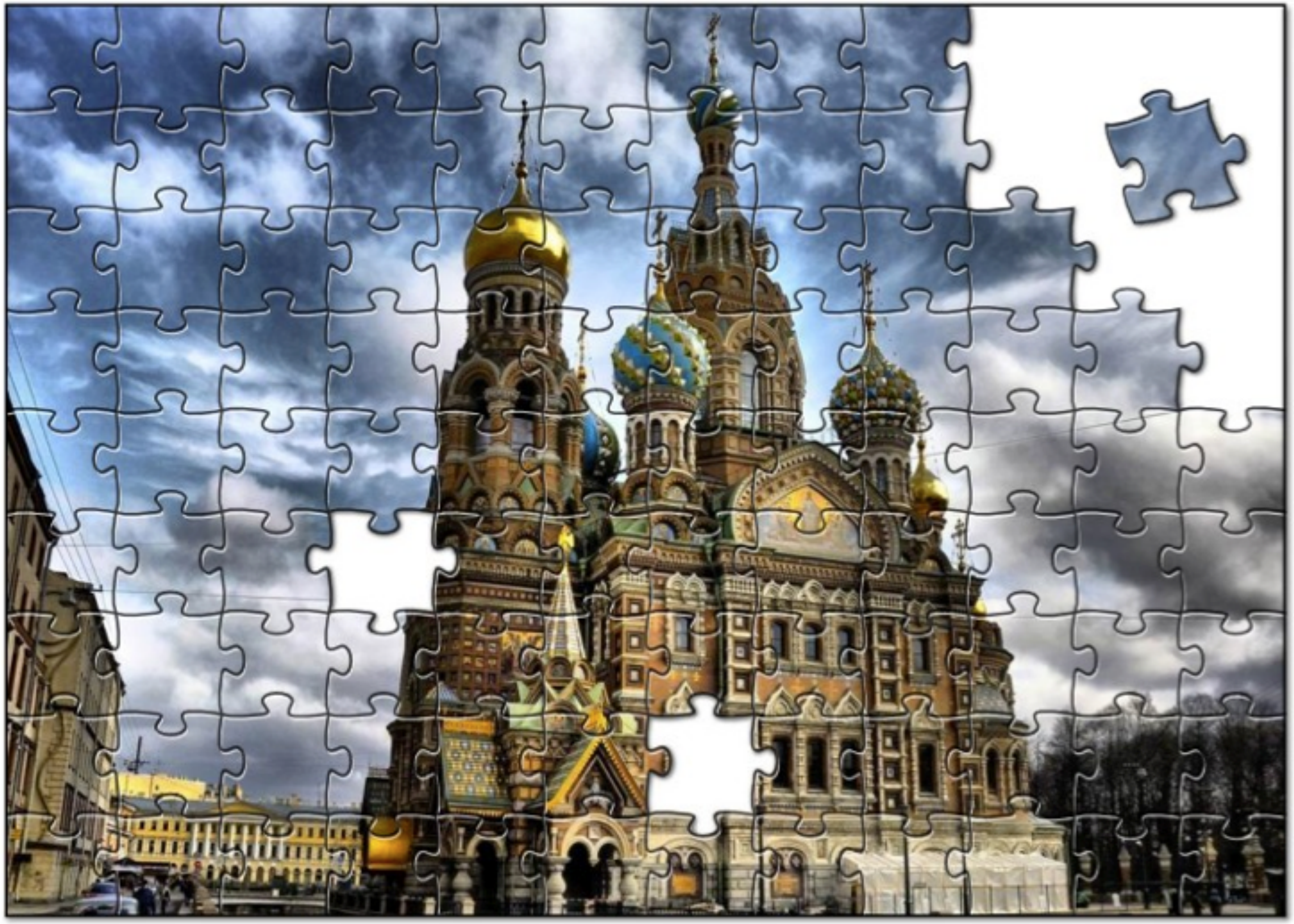
Expanding the SPAdes Toolbox

Anton Korobeynikov

Center for Algorithmic Biotechnology
St. Petersburg State University, St. Petersburg, Russia
http://bioinf.spbau.ru/spades



SPAdes (Saint Petersburg Assembler)

SPAdes

- Originally designed as single-cell assembler
- Can deal with highly uneven coverage and MDA-imposed chimeric reads
- Turned out to work well for multi-cell isolate assemblies
- One of two best assemblers in GAGE-B study by Salzberg’s lab (Magoc et al., Bioinf., 2013)
- The best bacterial genome assembler in the recent poll by acgt.me

SPAdes 3.5

- Improved memory consumption at the repeat resolution step (more than 2x)
- Integrated support for Lucigen NxSeq Long Mate Pair libraries
- Rewritten mismatch correction module
- Support for Oxford Nanopore reads for hybrid assemblies

Illumina + Nanopore Hybrid Assemblies

	Illumina only	**Ilmn + Nanopore**
Contigs > 500 bp	92	**1**
Largest Contig	285414	4649811
Total Length	4649811	4654532
Reference Length	4639675	4639675
NG50	133088	**4649811**
NG75	64475	4649811
Misassemblies*	0 (0)	6 (0)*
Genome fraction (%)	98.14	99.99

Illimina 2x100 bp E. coli K12 reads are available from http://bioinf.spbau.ru/spades
Nanopore reads from Nick Loman

* Misassemblies are not real, this is the difference wrt the reference

SPAdes 3.6

BayesHammer improvements:

- Removed 2^{32} k-mer limit (bigger genomes!)
- Reduced memory consumption (2x–4x)
- Much faster (e.g. 36h → 8h)
- Completely rewritten read correction procedure: faster and more precise

SPAdes improvements:

- Significantly reworked repeat resolution and scaffolding module

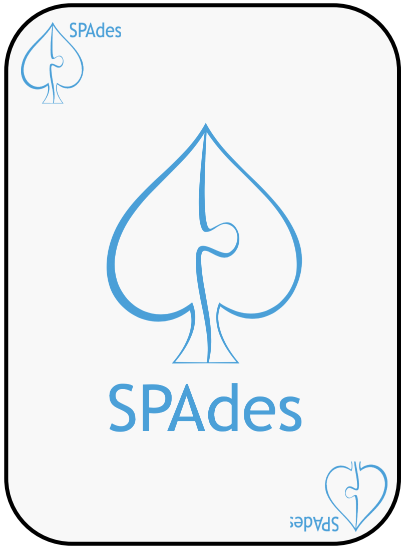
SPAdes Toolbox

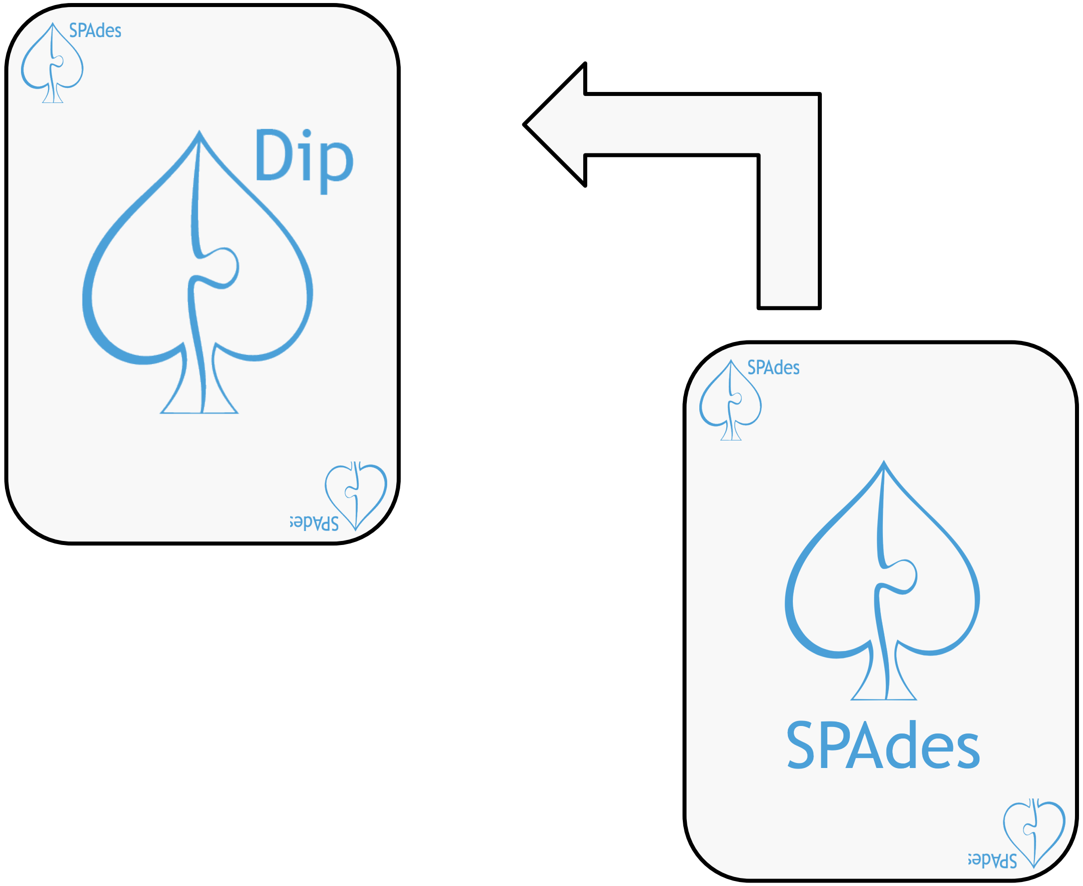
- Developed from the beginning as a set of modular and reusable parts
- Different “stages” of an assembler can be stacked together and share common information
- Allows one to assemble an assembler-like application from different building blocks

SPAdes Toolbox

- Developed from the beginning as a set of modular and reusable parts
- Different "stages" of an assembler can be stacked together and share common information
- Allows one to assemble an assembler-like application from different building blocks

And so we did!


SPAdes
SPAdes
SPAdes


SPAdes
Dip
SPAdes
SPAdes
SPAdes
SPAdes
SPAdes

dipSPAdes

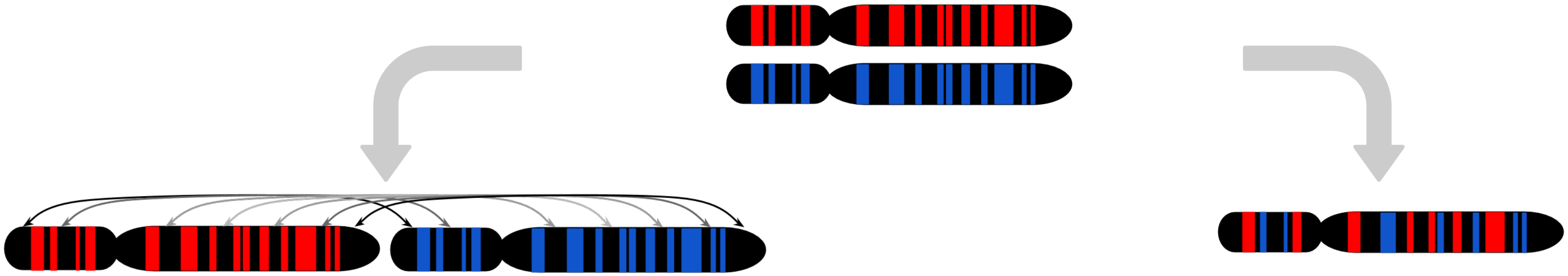
The first de Bruijn graph assembler designed for highly polymorphic diploid genomes:

Fungus
heterozygosity up to 20%

Sea squirts
heterozygosity up to 12%

Plants
avg heterozygosity 7%

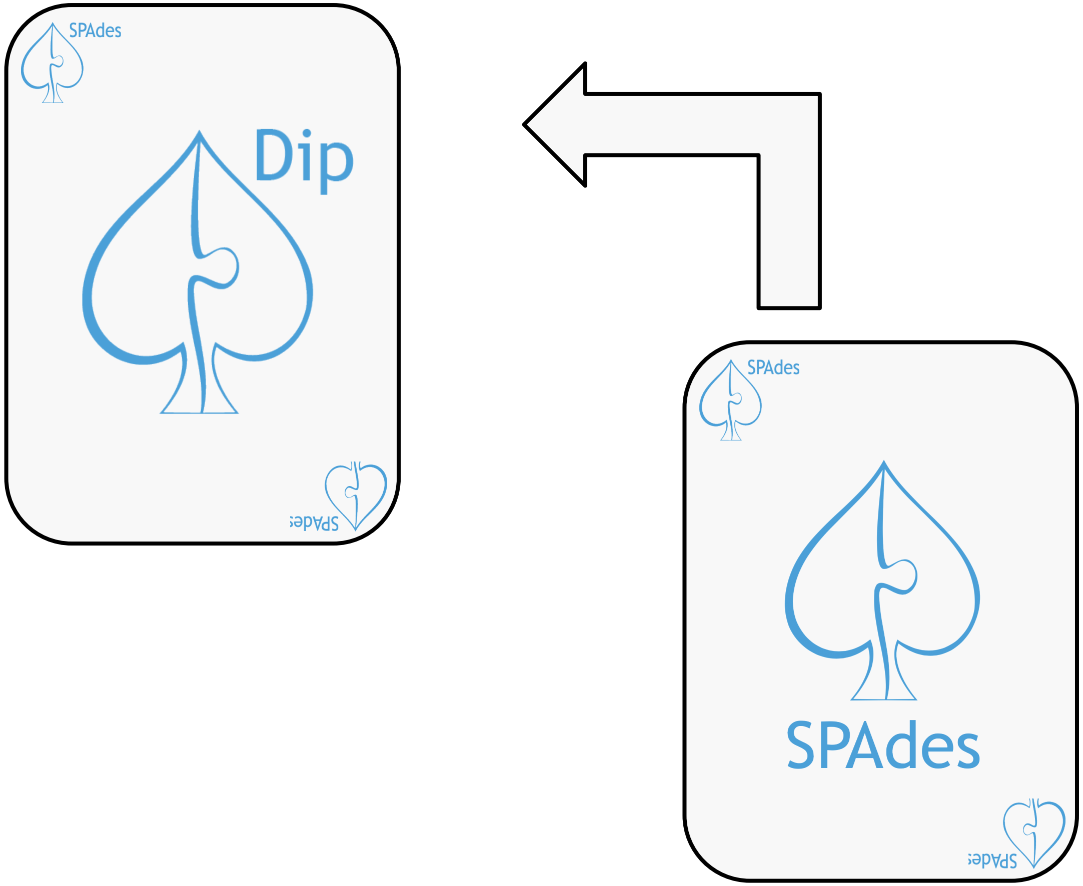
Insects
avg heterozygosity 9%



conventional approaches assemble such genome as two highly repetitive sequences and construct very fragmented assemblies

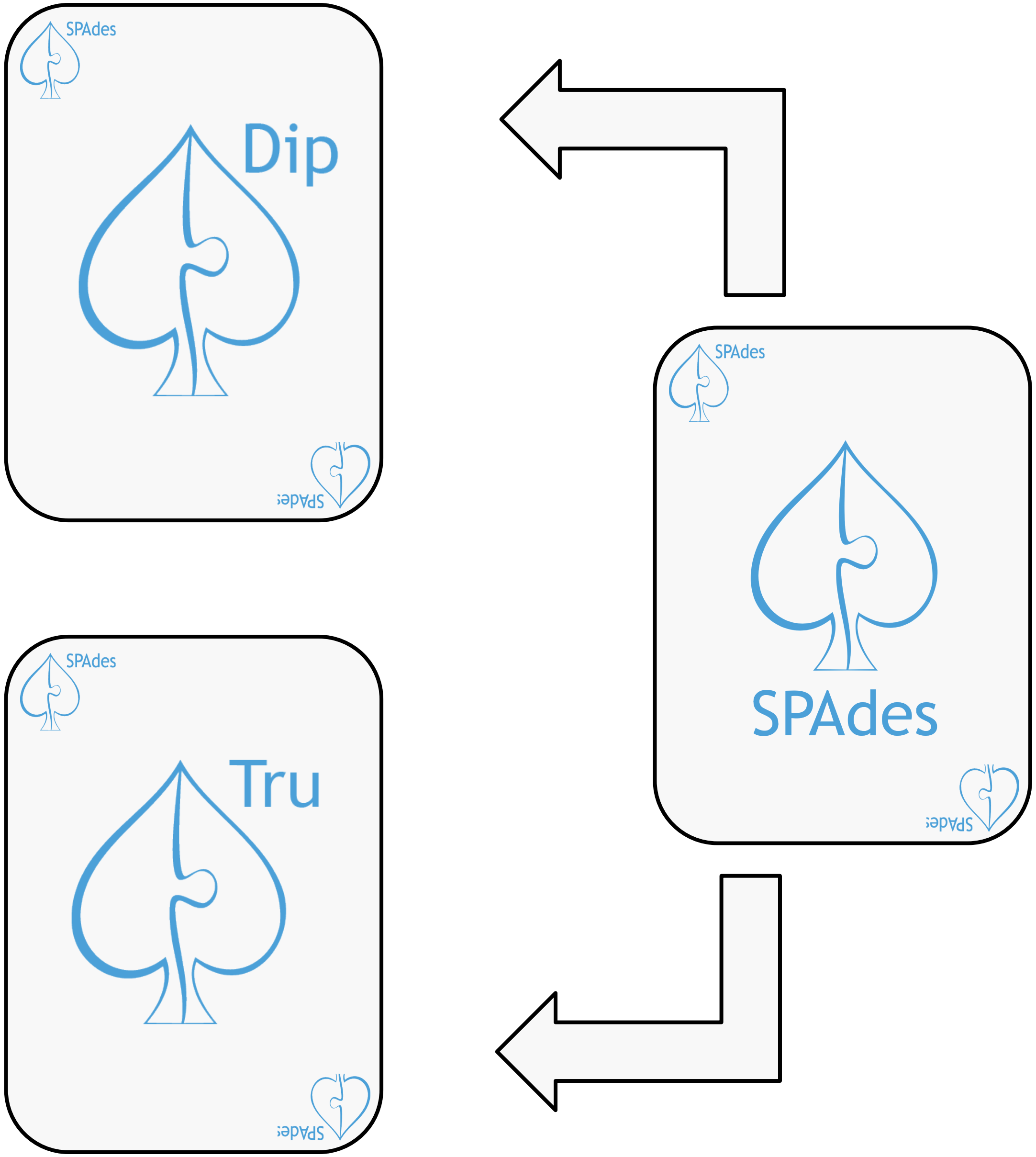
dipSPAdes constructs consensus for diploid haplomes and takes advantage of structure of de Bruijn graph for diploid genome to construct longer contigs

Yana Safonova, Anton Bankevich, Pavel A. Pevzner.
dipSPAdes: an assembler for highly polymorphic diploid genomes. J. of Comp. Biol., 2015

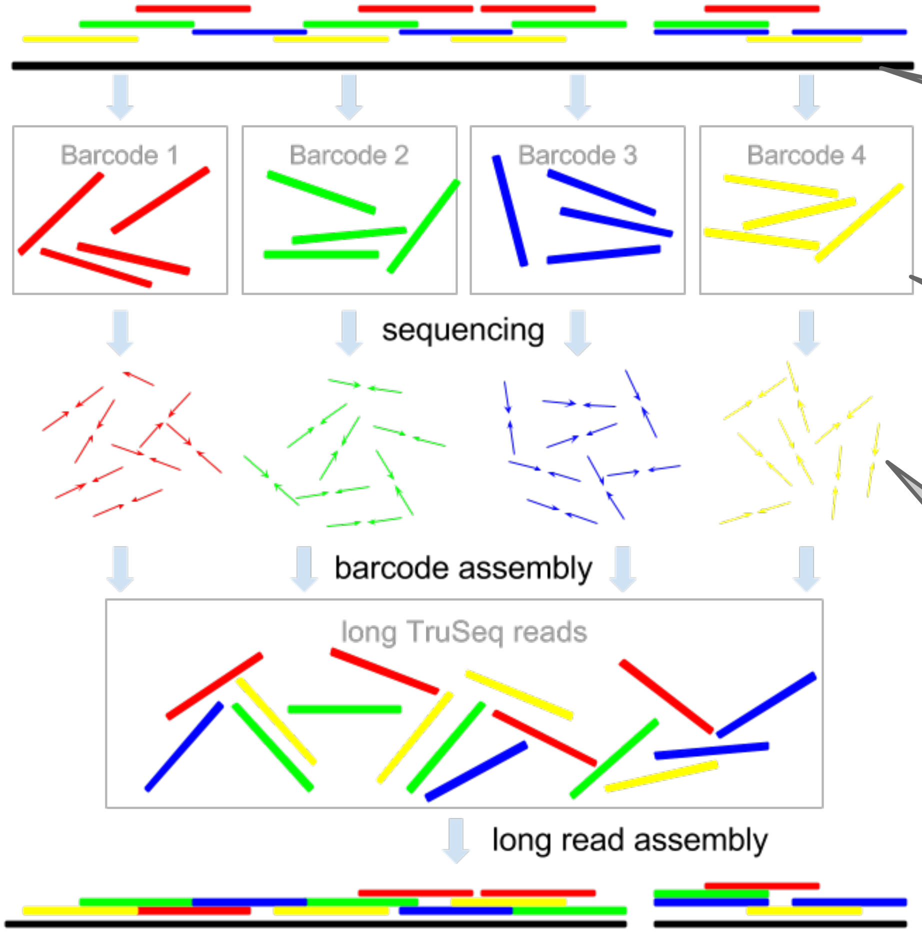


Dip

SPAdes


SPAdes
SPAdes
Dip
Tru

Illumina TruSeq

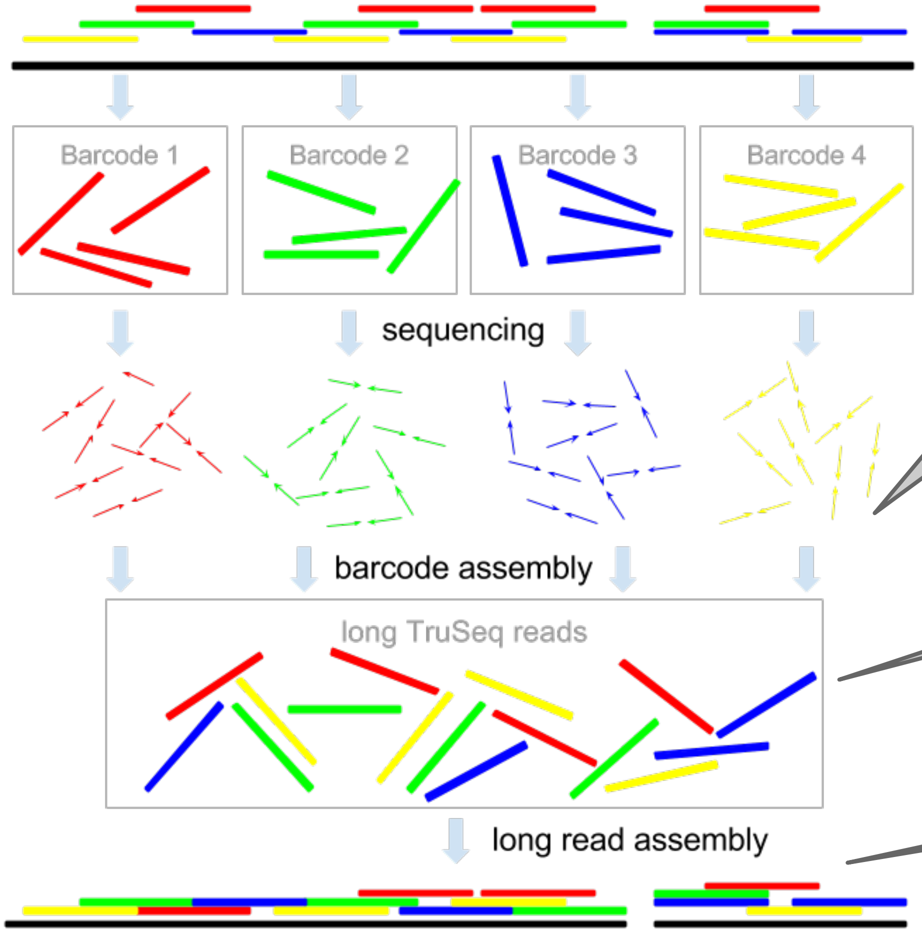


DNA is shred into 10Kb long fragments

Fragments are distributed among 96 pools

Pools are barcoded and sequenced by Illumina HiSeq

Illumina TruSeq



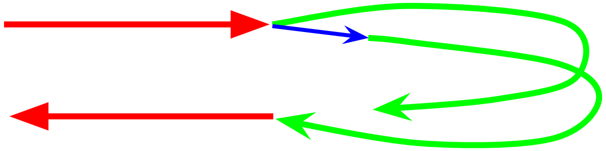
Reads from each pool are assembled separately

Resulting in virtual TruSeq long reads

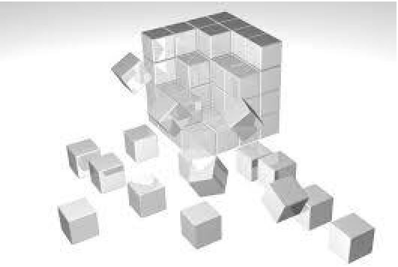
Which can be used as usual reads

Why SPAdes?

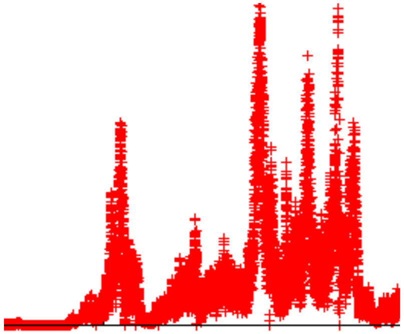
Complex repeat structure inherited from target genome



Interstrand chimeric connections

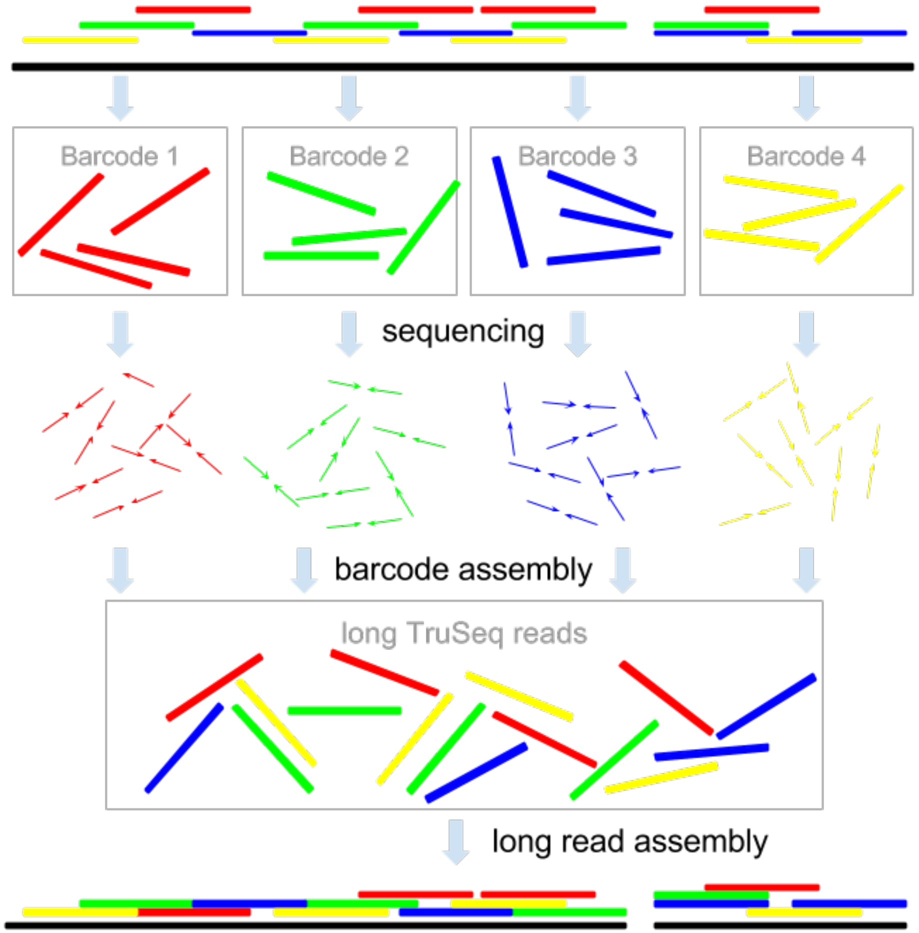


Fragmentation of barcode span



Uneven coverage

truSPAdes



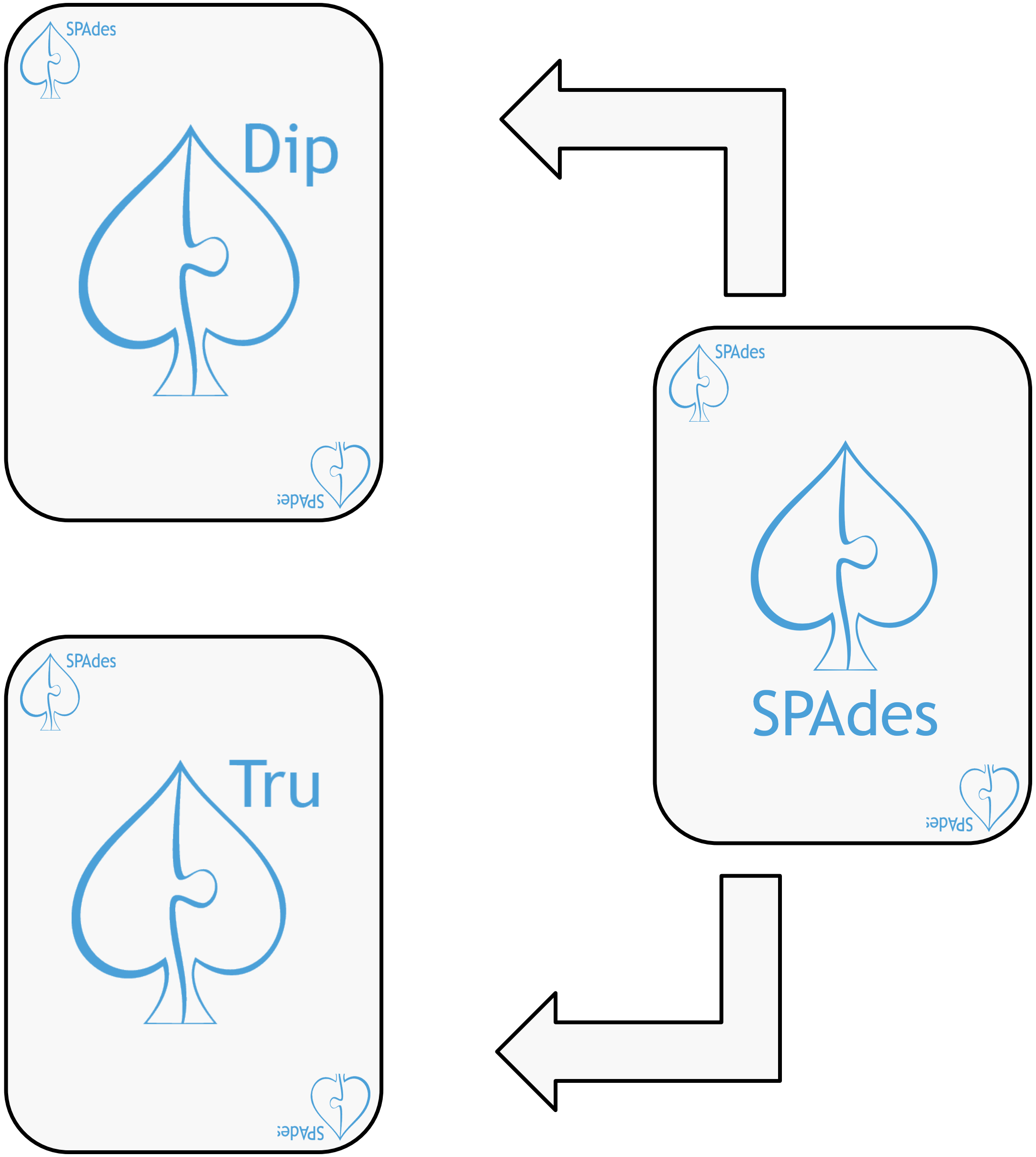
- SPAdes turned into assembler for pooled barcode data
- Tuning and refinements for TSLR data
- Accurate re-analysis of resulting contigs (virtual long reads) in order to reduce misassemblies

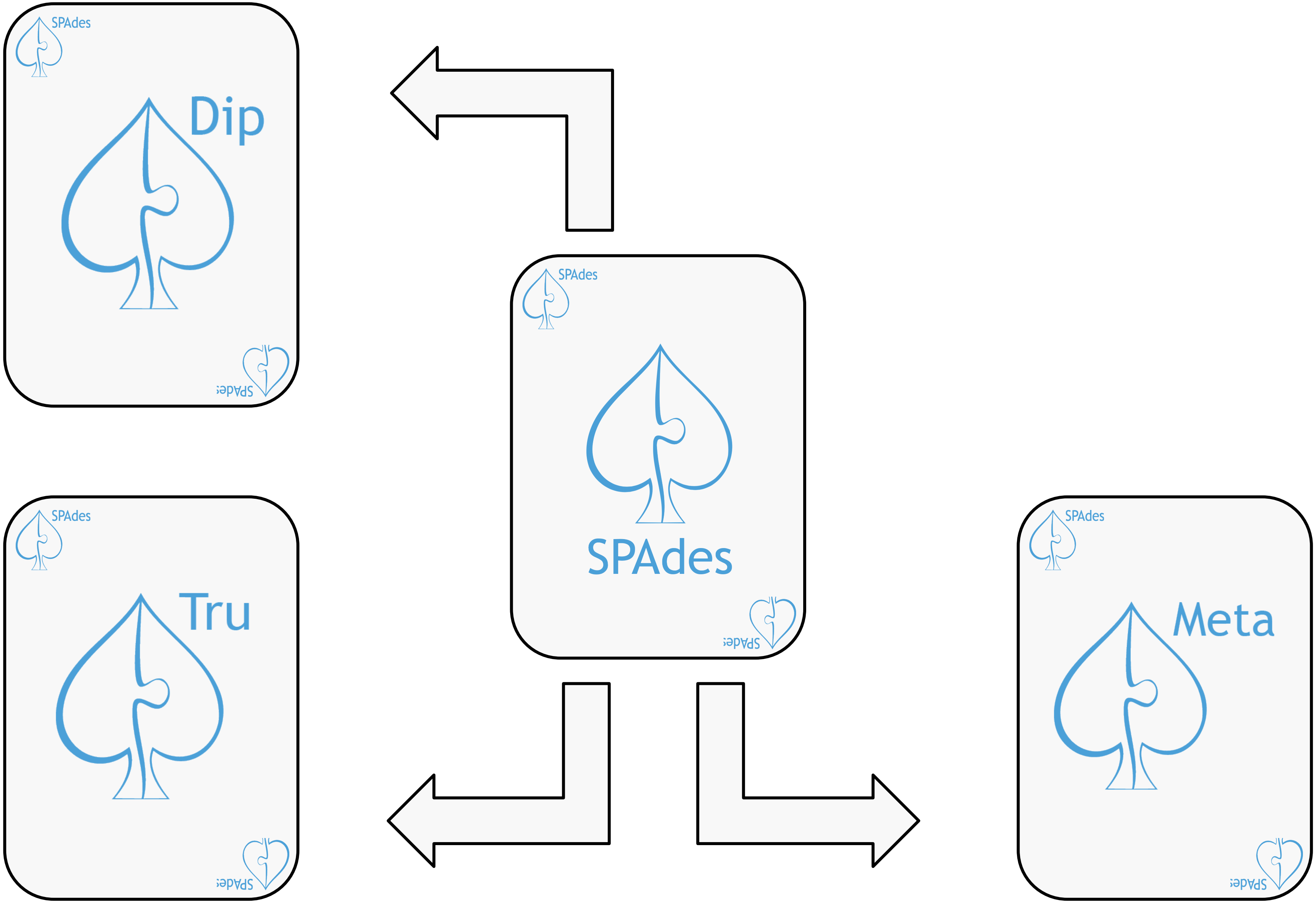
truSPAdes

	Illumina assembler	**Ray**	**SPAdes**	**truSPAdes**	**Ideal**
#contigs, pb*	419	414	677	430	≈300
#contigs (>8000 bp), pb	106	83	108	**126**	≈300
Total length (Mb), pb	2.2	1.8	**2.7**	2.3	≈3
N50	7 579	6 222	6 235	**8 250**	≈10 000
NGA50	5 235	2 511	4 770	**6 551**	≈10 000
#N's per 100 Kbp	0.9	3083	242	**0.3**	0
Misassemblies, pb	**1.8**	7	47	**3.1**	0
Mismatches per 100 Kbp	**75**	84	190	100	0

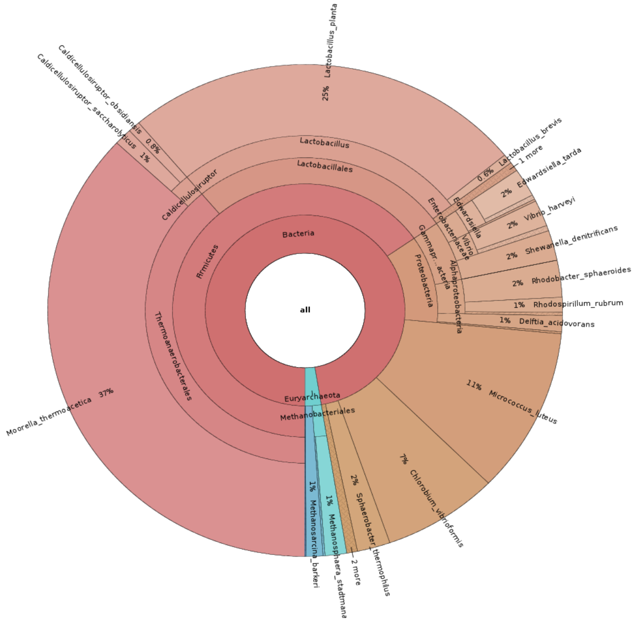
Human TSLR dataset

*pb - per barcode: average among all barcodes in dataset

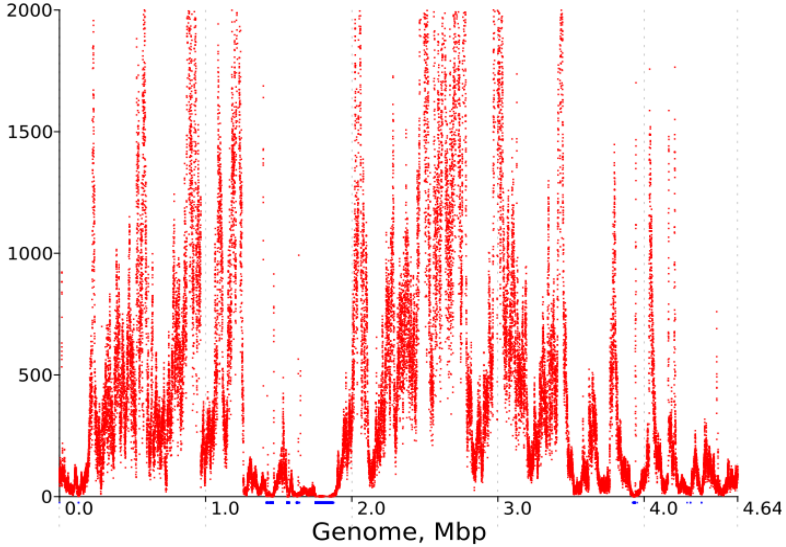

SPAdes
Dip
SPAdes
SPAdes
SPAdes
SPAdes
Tru
SPAdes


SPAdes
Dip
Tru
Meta

metaSPAdes



Relative abundance of species in metagenome

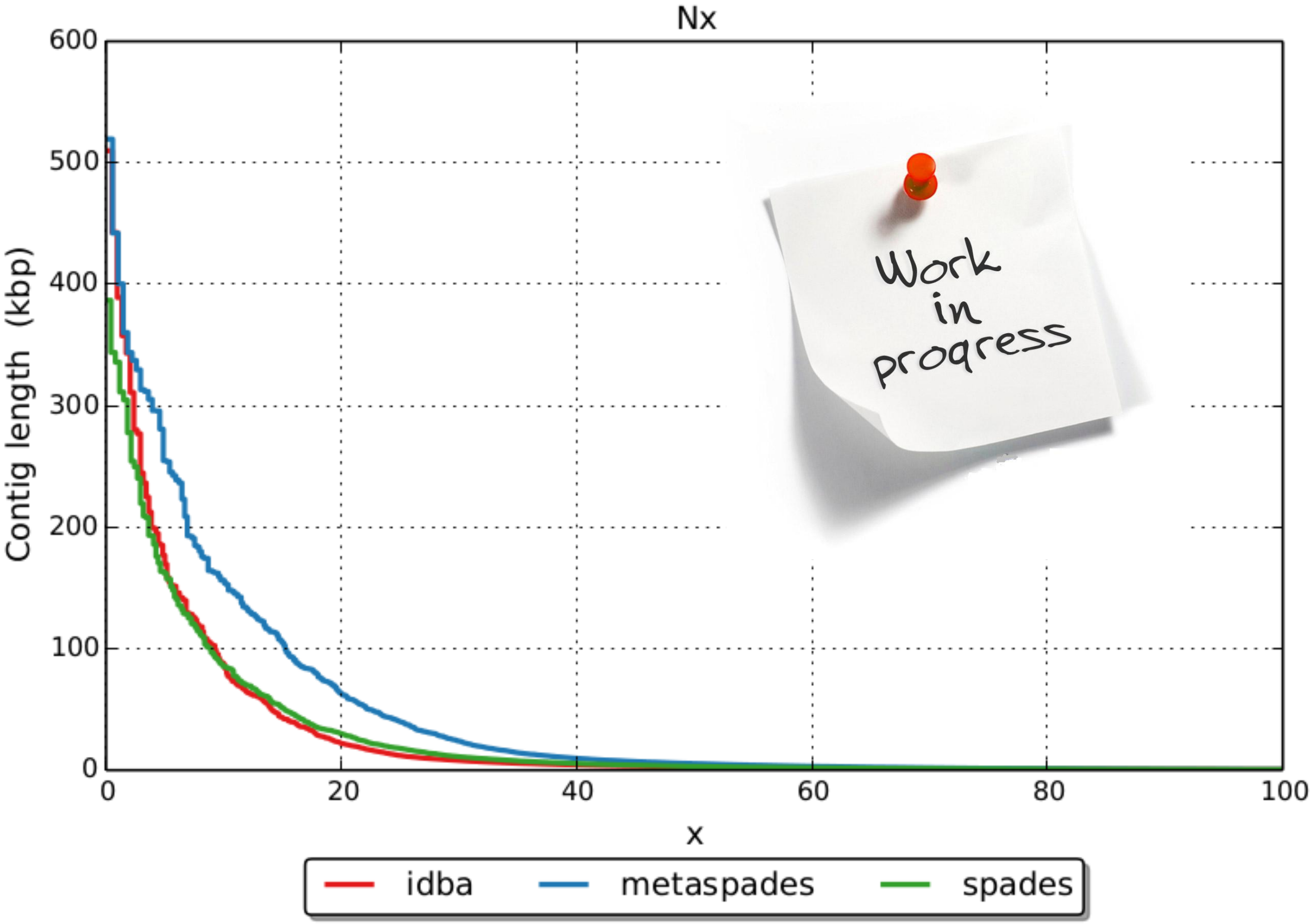


Coverage of single-cell *E. coli* sample

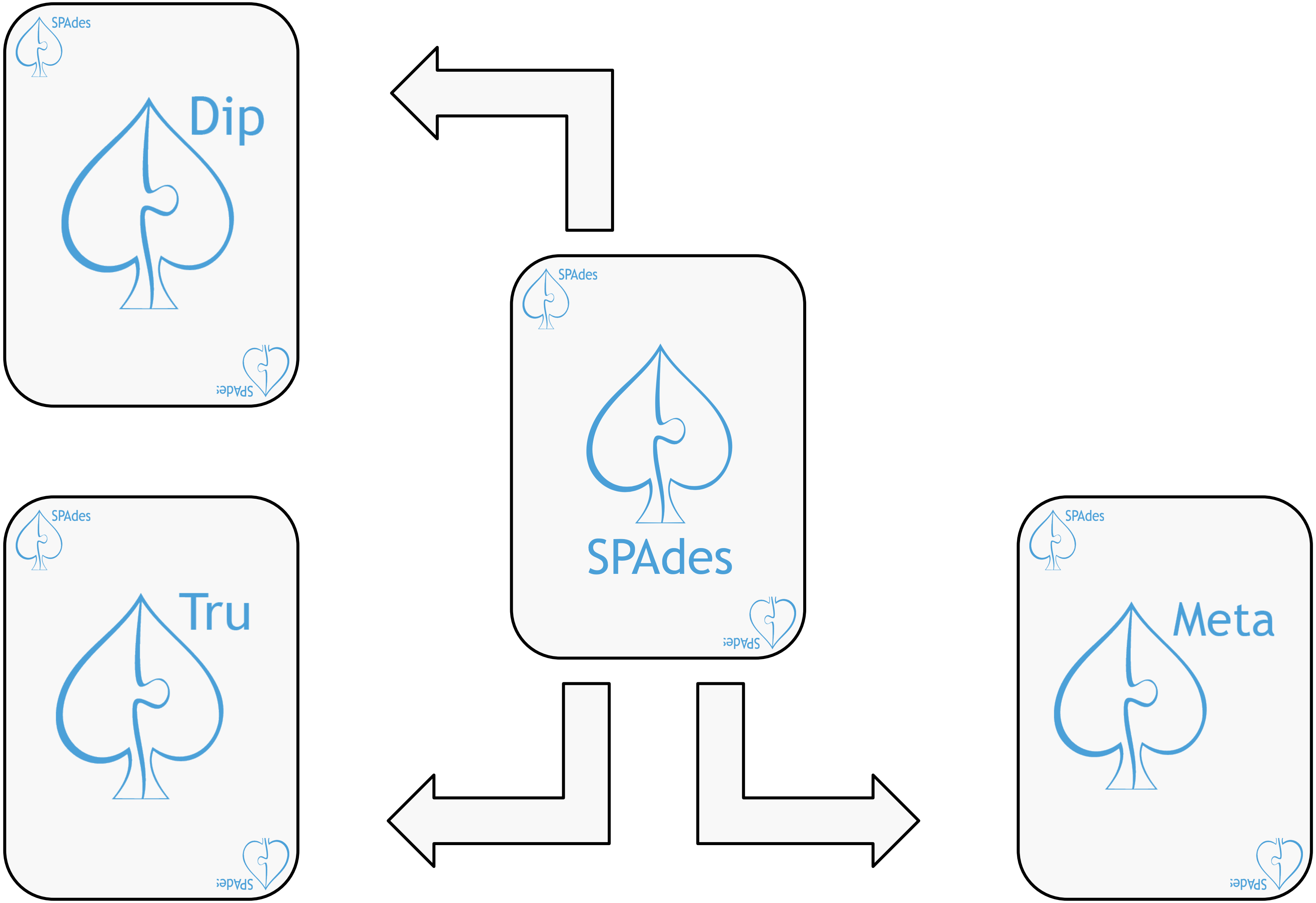
Genome assembly of species with extremely different abundances is similar to assembly of MDA data

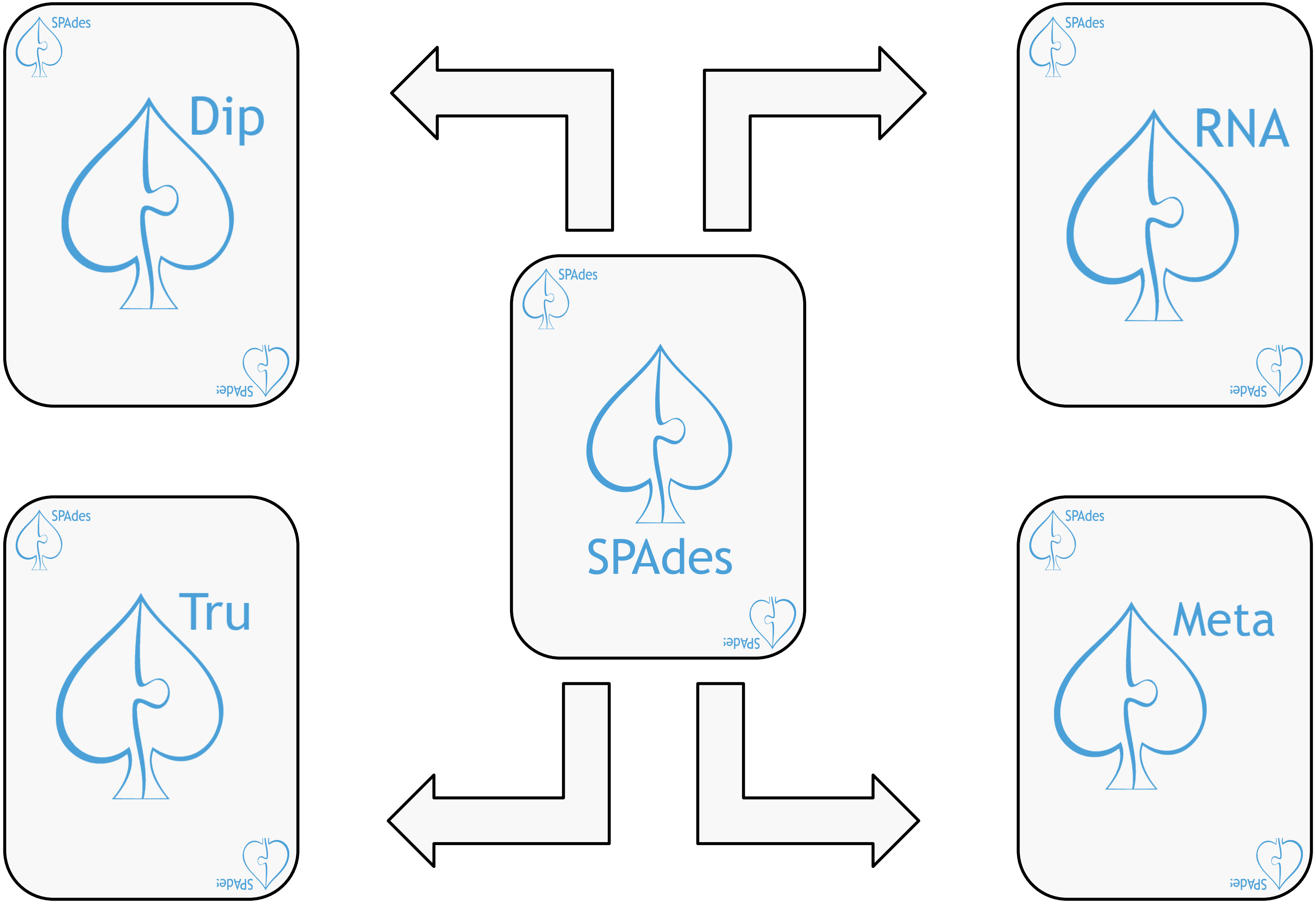
This is what SPAdes was designed for!

metaSPAdes



SRX024329 (HMP data) Nx plot


Dip
SPAdes
Tru
Meta


SPAdes
Dip
RNA
SPAdes
Tru
Meta

RNA-Seq assembly

- Trinity (Grabherr et al., Nat. Biotech., 2011)
- Oases (Schulz et al., Bioinf., 2012)

Who needs yet another RNA-Seq assembler?

RNA-Seq assembly

- Trinity (Grabherr et al., Nat. Biotech., 2011)
- Oases (Schulz et al., Bioinf., 2012)
- IDBA-tran (Peng et al., Bioinf., 2014)
- IDBA-MTP (Peng et al., RECOMB 2014)
- SOAPdenovo-Trans (Xie et al., Bioinf., 2014)
- StringTie (Pertea et al., Nat. Biotech., 2015)
- ….

Means there is a space for improving *de novo* transcriptome assemblers

How does a *single-cell genome* assembler perform on a transcriptome dataset?

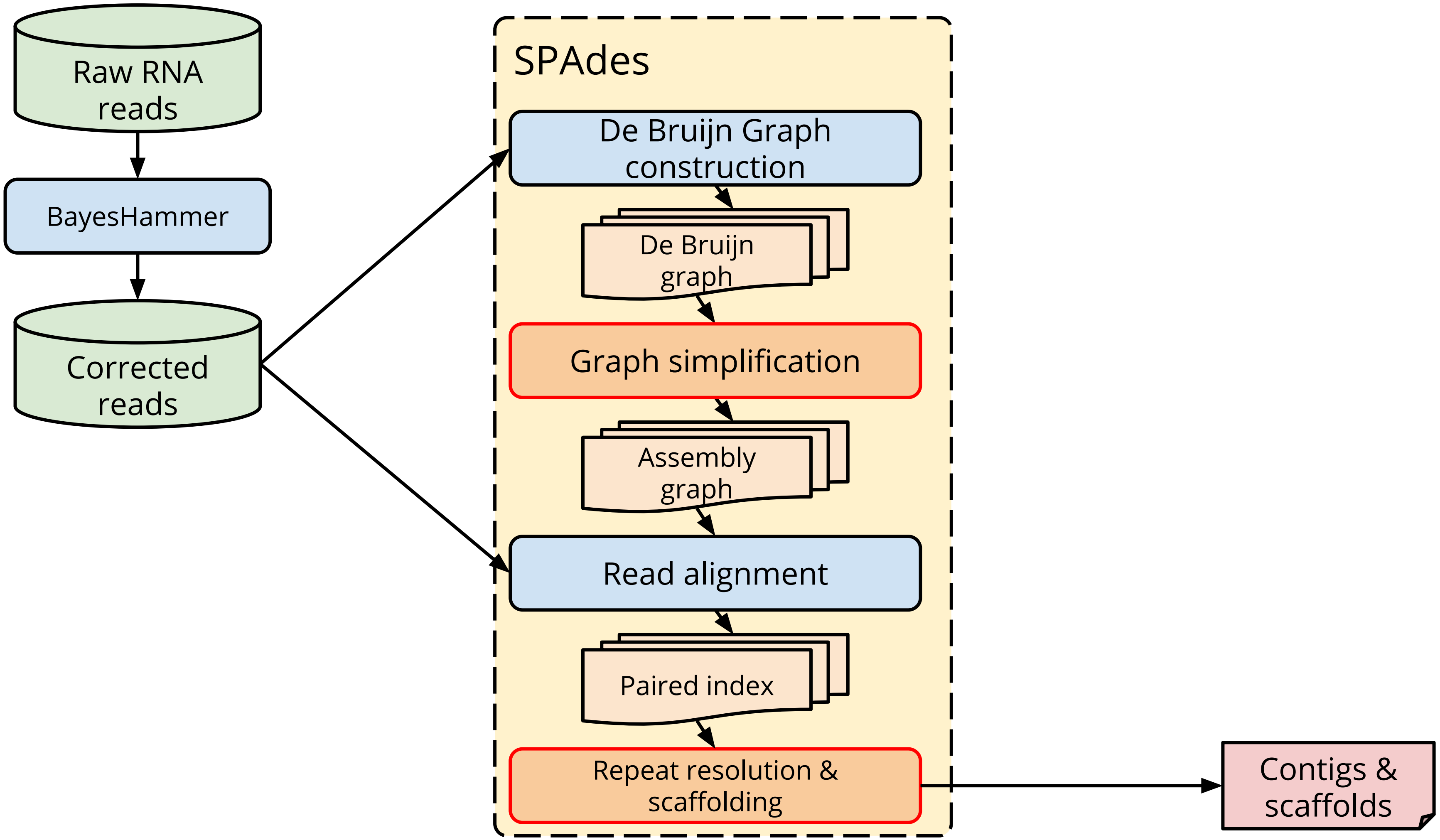
How does a *single-cell genome* assembler perform on a transcriptome dataset?

Quite well:

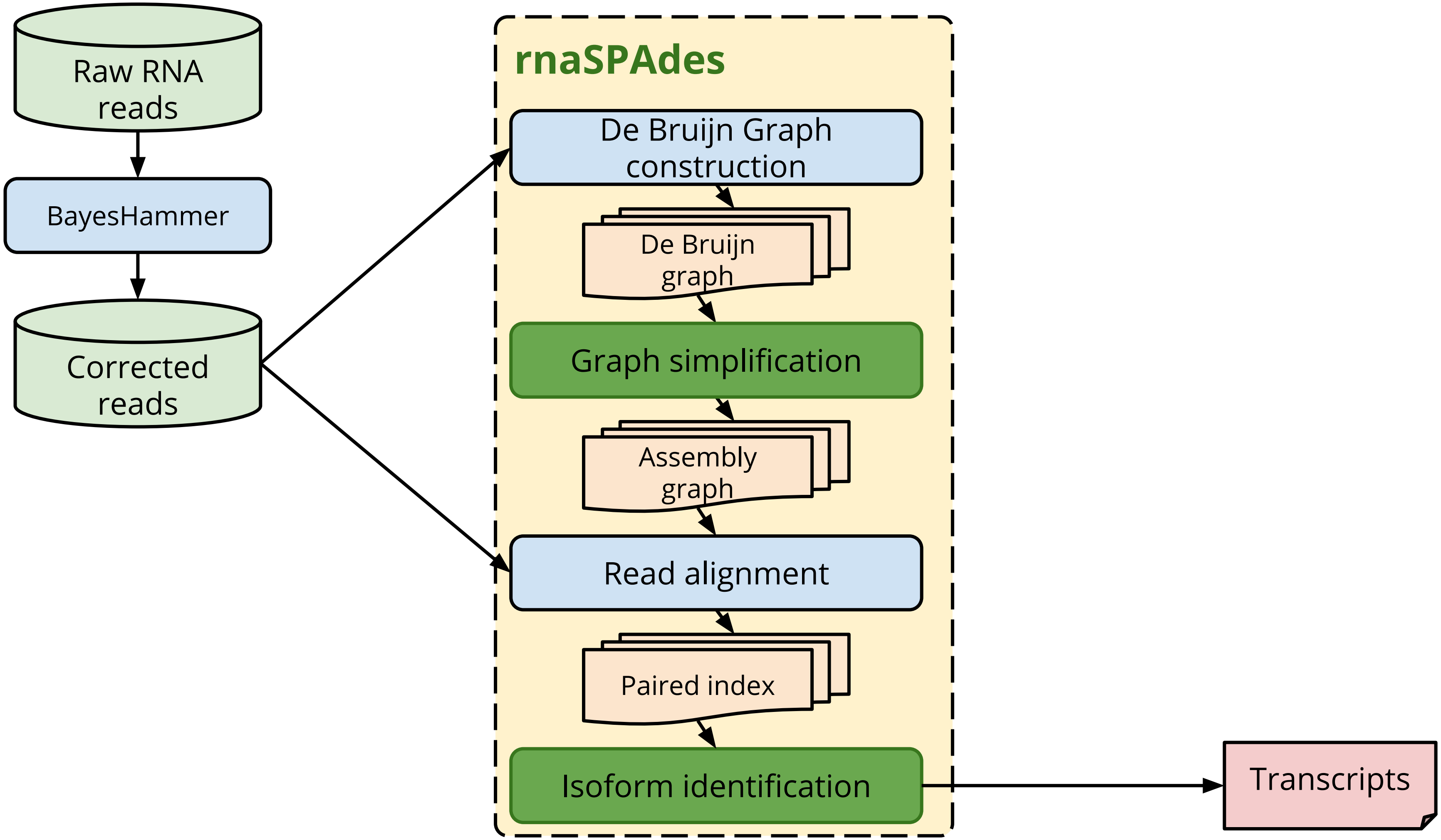
	IDBA-tran	**SOAPdenovoTrans**	**Trinity**	**SPAdes**
Transcripts	2872	2725	2171	3339
N50	312	213	309	**370**
Aligned	2845	2693	2150	**3230**
Unaligned	27	32	**21**	109
Avg. mismatches per transcript	0.447	0.456	**0.341**	0.57
Total annotation coverage	0.075	0.052	0.058	**0.1**
Partially-assembled isoforms (>30%)	886	582	713	**1119**
Fully-assembled isoforms (>90%)	96	53	91	**234**
Partially-annotated transcripts (>30%)	2611	2493	2009	**2967**
Fully-annotated transcripts (>90%)	1436	1449	1108	**1553**

Yeast RNA-Seq dataset

From SPAdes to rnaSPAdes



From SPAdes to rnaSPAdes



rnaQUAST

- One cannot develop an assembler without having an assembly quality assessment tool
- Based on our experience with SPAdes and QUAST, developing such a tool is not an easy task
- *Parallel* development of rnaSPAdes and rnaQUAST is crucial for the success of both tools
- rnaQUAST is tool for analysing assembled transcripts using various metrics (via the reference genome and / or genome annotation)

rnaSPAdes

	IDBA-tran	**SOAPdenovoTrans**	**Trinity**	**SPAdes**	**rnaSPAdes**
Transcripts	2872	2725	2171	3339	**6954**
N50	312	213	309	**370**	303
Aligned	2845	2693	2150	**3230**	**6692**
Unaligned	27	32	**21**	109	262
Avg. mismatches per transcript	0.447	0.456	**0.341**	0.57	**0.35**
Total annotation coverage	0.075	0.052	0.058	**0.1**	**0.105**
Partially-assembled isoforms (>30%)	886	582	713	**1119**	**1135**
Fully-assembled isoforms (>90%)	96	53	91	**234**	188
Partially-annotated transcripts (>30%)	2611	2493	2009	2967	**6138**
Fully-annotated transcripts (>90%)	1436	1449	1108	**1553**	**4094**

Yeast RNA-Seq dataset

When?

- SPAdes 3.6: end June
- dipSPAdes: included into SPAdes
- rnaSPAdes: beta mid June, EAP
- truSPAdes: beta mid summer
- metaSPAdes: beta end summer

Acknowledgement

SPAdes team:

Dmitry Antipov
Anton Bankevich
Elena Bushmanova
Alexey Gurevich
Dmitry Meleshko
Sergey Nurk
Andrei Przhibelski
Yana Safonova

Alla Lapidus
Pavel Pevzner

This work was supported by the Russian Science Foundation (grant #14-50-00069)

Thank you!

SPAdes

http://bioinf.spbau.ru/spades